Sentence by sentence Sentiment Extraction

This research deals with sentence by sentence sentiment extraction and the aim is to show the strengts of this approach while underlying possible improvements.

Data

Data used are *wheat_articles* articles from Michael Kao's *keyword_extraction.py* script: they are 15017 wheat related articles taken from the original 60937 articles corpus.
From these articles just the sentences that contain the word *"wheat"* are extracted and they are then analyzed using VADER analysis tool.
The outcome is a 3 columns dataframe containing: analyzed sentence, compound sentiment score and original article date and time.
The dataframe, in python called *"sentiment"*, is then written in a *.csv* file called "*df.csv*".
The method still has some flaws, for example sentences that are too short to be properly analyzed (therefore resulting in NA), badly formatted sentences from badly formatted articles and repeated sentences from duplicated articles. However it has to be underlined that the main source of all these issues is badly formatted articles, therefore these problems will be solved by cleaning *ex-ante* the corpus of wheat-related articles.

Results

A few sentences will be reviewed in order to stress the strenghts and improvement margins of this approach.
From the original data 163307 sentences were chosen from 15017 wheat-related articles.
It has to be underlined that 40% sentences were rated 0, namely they bear no sentiment because of issues that will be discussed later.
In the comments it's assumed that positive sentiment is related to price increases, while negative sentiment is related to price decreases.

Phrase ID	**Sentence**	**Compound Sentiment**
1)	*Wheat: The wheat market closed around 4-6 cents higher on the day and with gains of varying magnitude for the week - led by a Minneapolis market searching for quality.*	*0.34*
2)	*Indian temperatures 2-4C above normal are delaying wheat plantings and could ultimately cut yields.*	*-0.2732*
3)	*The projected glut in 2015/16 global wheat supplies is pressuring export prices to the lowest level since July 2010.*	*-0.6124*
4)	*Jan 16 London wheat was up GBP0.15/tonne at GBP112.70/tonne.*	*0*
5)	*Italy represents only a small percentage of total U.S. wheat sales.*	*0*
6)	*Jul 15 London wheat was GBP1.50/tonne lower at GBP111.50/tonne.*	-0.296

Phrase 1 talks about wheat price and is quite *verbose*: given this sentence peculiarity VADER analysis tool correctly specifies a mildly positive compound sentiment.
Phrase 2 deals with weather delaying wheat production: also in this case VADER correctly specifies a negative compound sentiment given sentence *verbosity*. Moreover, thinking to wheat price correlation, influences on prices by different wheat-related topics could be assumed.

Phrase 3 deals in a *verbose* way with prices: in this case VADER correctly specifies a strongly negative compound sentiment, due to the use of the verb *"pressuring"* and the superlative *"lowest"*.
Phrase 4 is one of the uncorrectly analyzed sentences: VADER reports 0 compound sentiment, but the fact that price is going up bears economic sentiment. This result is given by the fact that the sentence is not *verbose*: it's a 100% neutral economic statement. Unfortunately in many articles there are sentences of this type, therefore part of resulting zeros is given by them. In particular the resulti s given by the absence of the word *"up"* in VADER lexicon.
Phrase 5 is a correctly specified no sentiment bearing sentence: the fact that Italy imports a small percentage of US wheat total sales logically has no direct correlation and no predictive power in respect to prices.
Phrase 6 it's similiar to *phrase 4*, but it's correctly analyzed by VADER: probably the difference is given by the presence of the adjective *"lower"* that, being present in VADER lexicon, gives the negative result.

Conclusion:
If the aim is to capture wheat related sentiment, sentence by sentence sentiment extraction approach gives sharp results, even using VADER sentiment analysis tool.
On one hand phrases 1, 2, 3 and 6 are in fact correctly specified due to their *verbosity*, while on the other hand phrase 4 is uncorrectly specified given the absence of *human-sentiment*, namely lack of emotional statements, neutral style and no *verbosity*.
Since in the dataframe the number of 0 compound sentiment sentences is pretty high and given the fact that *economic-sentiment* bearing sentences (phrase 4) must be discriminated from no sentiment bearing sentences (phrase 5), it's advisable to work on the dictionary by improving its lexicon.
Moreover some more work should be spent on articles clean in order to remove duplicated articles (or duplicated parts of them) and badly formatted articles.
In conclusion it can be affirmed that a sentiment index can be created by cleaning the articles, selecting only the sentences related to wheat, extracting sentiment from them with a refined dictionary and filtering sentiment scores using the Kalman filter.
Hopefully this sentiment index should be correlated with wheat price series.

<u>Further Research Topics</u>

- Articles cleaning.
  Many articles are duplicated or contain duplicated parts (the ones from Agrimoney for example).

- Dictionary refinement.
  As already reported in the Proof Of Concept analysis performed in dictionaries review, VADER (as other dictionaries) is not able to capture *economic-sentiment*, namely it gives no sentiment to neutral economic statements that may contain some predictive power towards prices. It's therefore advised to correct the lexicon used.

- (Conceptual) Which sentence types should be taken into account?
  Both a sentence describing increasing prices and a sentence where increased production is reported are positive sentiment bearing, but in the economic theory the effect on prices is the opposite. Should only strictly price related sentences be taken into account or should a more wide interpretation of sentiment be taken into account?

- (Technical) Link to The Reading Machine.
  If this approach will be selected, a script that is directly linked to The Reading Machine has to be written.